

A random algorithm for low-rank decomposition of large-scale matrices with missing entries

Yiguang Liu



Abstract—A Random SubMatrix method (**RSM**) is proposed to calculate the low-rank decomposition $\hat{\mathbf{U}}_{m \times r} \hat{\mathbf{V}}_{n \times r}^T$ ($r < m, n$) of the matrix $\mathbf{Y} \in R^{m \times n}$ (assuming $m > n$ generally) with known entry percentage $0 < \rho < 1$. **RSM** is very fast as only almost $\mathcal{O}(mr^2\rho^r)$ or $\mathcal{O}(n^3\rho^{3r})$ floating-point operations (flops) are required, compared favorably with $\mathcal{O}(mnr + r^2(m + n))$ flops required by the state-of-the-art algorithms. Meanwhile **RSM** is very memory-saving as only $\max(n^2, mr + nr)$ real values need to save. With known entries homogeneously distributed in \mathbf{Y} , sub-matrices formed by known entries are randomly selected from \mathbf{Y} with statistical size $k \times n\rho^k$ or $m\rho^l \times l$ where k or l takes $r + 1$ usually. According to the just proved theorem that noises are less easy to cause the subspace related to smaller singular values to change with the space associated to anyone of the r largest singular values, the $n\rho^k - k$ null vectors or the $l - r$ right singular vectors associated with the minor singular values are calculated for each submatrix. The vectors are null vectors of the submatrix formed by the chosen $n\rho^k$ or l columns of the ground truth of $\hat{\mathbf{V}}^T$. If enough sub-matrices are randomly chosen, $\hat{\mathbf{V}}$ and accordingly $\hat{\mathbf{U}}$ are estimated. The experimental results on random synthetical matrices with sizes such as 131072×1024 and on real data sets such as dinosaur indicate that, **RSM** is $4.97 \sim 88.60$ times faster than the state of the art. It, meanwhile, has considerable high precision achieving or approximating to the best.

Index Terms—Low-rank matrix decomposition, random submatrix, complexity, memory-saving

1 INTRODUCTION

Low rank approximation of matrices with missing entries is ubiquitous in many areas such as computer vision, scientific computing and data analysis, etc. How to quickly implement low-rank approximation of large-scale matrices with perturbed and missing entries is significant due to the challenge of big data usually with many entries or feature components missed or perturbed. Here we restrict our attention to using random techniques to efficiently find the low-rank structure, and for the latest advancement in this direction we refer the reader to refs. [1], [2]. In [1] random algorithms are introduced to reduce matrix sizes and then the low-rank decomposition is manipulated deterministically on the reduced matrix. Ref [2] thoroughly reviewed the probabilistic algorithms for constructing the low-rank approximation of a given matrix, but the algorithms are only applicable to matrices without missing entries.

Up to now, low-rank decomposition of the matrices with missing entries usually performs in deterministic ways. Though space constraints preclude us from reviewing the extensive literature on the subject, and even it is impossible to comment each algorithm [3]–[15] in a little detail. However, we still feel it is necessary to point out some milestone algorithms in this field. The deterministic way [3] proposed in 2013 has the state-of-the-art low-rank approximating capability. Its flops are typically proportional to

$$C_r = r^2 \log(r) + mnr + r^2(m + n) \quad (1)$$

for solving the low-rank decomposition problem

$$\min_{\mathbf{U}, \mathbf{V}} \|\mathbf{W}_{m \times n} \odot (\mathbf{Y}_{m \times n} - \mathbf{U}_{m \times r} \mathbf{V}_{n \times r}^T)\| \quad (2)$$

where $r < \min(m, n)$. The operator $\|\cdot\|$ denotes some norm such as L_1 , L_2 or Frobenius norm, and \odot the Hadamard multiplication operator. The entry of \mathbf{W} , w_{ij} , takes 1 if the corresponding entry in \mathbf{Y} , y_{ij} , is known, otherwise 0. In each iteration, C_r flops are required, and how many iterations are required depends on precision requirements and the algorithm's own convergence capability. The index ρ showing the percentage of known entries in \mathbf{Y} is defined below:

$$\rho = \frac{1}{mn} \sum_{i,j} w_{ij}. \quad (3)$$

OptSpace [16] is a method based on optimization over the Grassmann manifold with a theoretical performance guarantee for the noiseless case. L2-wiberg [17] has very high global convergence rate and is insensitive to initialization for a wide range of missing data entries. However, the breakdown point of L2-Wiberg is not at the theoretical minimum due to the lack of regularization, as indicated in [3]. Meanwhile, L2-Wiberg is memory-consuming as two dense matrices with sizes $mnp \times mr$ and $mnp \times nr$ are required. The storage requirement makes L2-Wiberg not applicable to large-scale matrices, and this phenomenon is confirmed later in Section 5.

Our scheme **RSM** needs to randomly choose known entries from \mathbf{Y} so as to form submatrices, and its CPU time is almost proportional to

$$C_{\text{RSM}} = \mathcal{O}(mr^2\rho^r) \text{ or } \mathcal{O}(n^3\rho^{3r}) \quad (4)$$

• Yiguang Liu is with Vision and Image Processing Laboratory, College of Computer Science, Sichuan University, Chengdu, Sichuan Province, China, 610064. E-mail: liuyg@scu.edu.cn

mainly spent in calculating some singular vectors of submatrices each. Comparing (4) with (1) demonstrates the efficiency superiority of **RSM**.

Apart from the efficiency advantage, the randomized approach is more robust and can easily be reorganized to exploit multiprocessor architectures when compared to deterministic ways in solving (2) [2]. Besides, deterministic ways need to save at least $mn + 2r(m+n)$ values while the proposed method only needs memory space to save $\max(n^2, mr + nr)$ real numbers. The method proposed in this paper can be seen as an extension of the methods introduced in refs. [1], [2] where randomized techniques are used to perform the low-rank decomposition of large-scale matrices without missing entries or to reduce matrix size. In contrast, **RSM** directly applies random schemes to low-rank decomposition, preserving both efficiency and robustness traits of random techniques.

This article has the following structure: Section 2 sets the notation. Section 3 provides the relevant mathematical apparatus, followed by the algorithm description in Section 4. Section 5 illustrates the performance of the proposed algorithm via simulations on the synthetic and the real data sets. The conclusions are drawn in Section 6 with future works proposed therein.

2 NOTATION

Here we set notational conventions employed throughout the paper. Let

$$\mathbf{Y} = \bar{\mathbf{Y}} + \Psi \quad (5)$$

where $\Psi = [\psi_{ij}]$ denotes noise matrix, and the rank- r matrix $\bar{\mathbf{Y}} = [\bar{y}_{ij}]$ is the ground truth of \mathbf{Y} . SVD of $\bar{\mathbf{Y}}$ is denoted as $\bar{\mathbf{Y}} = \bar{\mathbf{U}}\bar{\Sigma}\bar{\mathbf{V}}^T$ where $\bar{\mathbf{U}} = [\bar{u}_1, \dots, \bar{u}_r] \in R^{m \times r}$, $\bar{\Sigma} = \text{diag}(\bar{\sigma}_1, \dots, \bar{\sigma}_r)$ and $\bar{\mathbf{V}} = [\bar{v}_1, \dots, \bar{v}_r] \in R^{n \times r}$. Similarly, u_i , σ_i and v_i represent the SVD of \mathbf{Y} , and the space spanned by vectors v_i, \dots, v_{i+r} is denoted as $\mathcal{S}_{v_i, \dots, v_{i+r}}$. Let $[n]$ denote the list composed of $1, 2, \dots, n$, $\langle k \rangle$ a list having k different natural numbers, and further $\mathbf{Y}_{\langle k \rangle \times \langle l \rangle}$ denote a sub-matrix of \mathbf{Y} formed by the entries at the intersections of $\langle k \rangle \subset [m]$ rows and $\langle l \rangle \subset [n]$ columns. We use $\text{vec}(A)$ to denote the column vector stacking the columns of the matrix A on top of one another, and assume $m \geq n$ without loss of generality across the article.

3 MATHEMATICAL APPARATUS

In (2), if \mathbf{V} becomes known, \mathbf{U} is accordingly solved. So in what follows we shall work exclusively with how to work out \mathbf{V} . The ground truth matrix $\bar{\mathbf{Y}}$ is unknown and so are \bar{v}_i for $1 \leq i \leq n$. The singular values associated with $\bar{v}_{r+1}, \dots, \bar{v}_n$ are all zero and are in equivalent importance, thus we do not care each concrete vector of $\bar{v}_{r+1}, \dots, \bar{v}_n$, and what we want to know is the space $\mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$. The space $\mathcal{S}_{v_{r+1}, \dots, v_n}$ is fixed by \mathbf{Y} , and the noise matrix Ψ makes $\mathcal{S}_{v_{r+1}, \dots, v_n}$ deviate from its ground truth $\mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$. How does Ψ affect the deviation can refer to the following two conclusions.

Theorem 1: For $1 \leq i \leq r$ and $r < j \leq n$, if

$$\|\sigma_i u_i^T - v_i^T \Psi^T\| \geq \|\sigma_j u_j^T - v_j^T \Psi^T\|, \quad (6)$$

then $\mathcal{S}_{v_{r+1}, \dots, v_n} = \mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$ absolutely holds.

Proof: By projecting the column vectors of $\bar{\mathbf{Y}}^T = \mathbf{Y}^T - \Psi^T$ on to v_1, \dots, v_n , we get the energy on v_i as

$$\|v_i^T (\mathbf{Y}^T - \Psi^T)\| = \|\sigma_i u_i^T - v_i^T \Psi^T\|.$$

For any unit vector $\kappa \in \mathcal{S}_{v_1, \dots, v_r}$, we have

$$\kappa = \sum_{i=1}^r \alpha_i v_i, \quad \text{s.t.} \quad \sum_{i=1}^r \alpha_i^2 = 1 \quad (7)$$

as v_1, \dots, v_r are unit vectors and are orthogonal to each other. The energy of projecting the column vector of $\bar{\mathbf{Y}}^T$ onto κ is

$$E_\kappa = \sqrt{\sum_{i=1}^r \alpha_i^2 \|\sigma_i u_i^T - v_i^T \Psi^T\|^2}$$

which combined with (7) indicates that

$$\min_{1 \leq i \leq r} \|\sigma_i u_i^T - v_i^T \Psi^T\| \leq E_\kappa \leq \max_{1 \leq i \leq r} \|\sigma_i u_i^T - v_i^T \Psi^T\|. \quad (8)$$

Similarly for any unit vector $\iota \in \mathcal{S}_{v_{r+1}, \dots, v_n}$, we have

$$\min_{r+1 \leq i \leq n} \|\sigma_i u_i^T - v_i^T \Psi^T\| \leq E_\iota \leq \max_{r+1 \leq i \leq n} \|\sigma_i u_i^T - v_i^T \Psi^T\|. \quad (9)$$

The two unit vectors κ and ι are randomly chosen in $\mathcal{S}_{v_1, \dots, v_r}$ and $\mathcal{S}_{v_{r+1}, \dots, v_n}$ respectively. By combining (6), (8) and (9) together, we can conclude that the energy gotten by projecting all column vectors of $\bar{\mathbf{Y}}^T$ onto any unit vector in $\mathcal{S}_{v_1, \dots, v_r}$ is larger than the energy projected onto any unit vector in $\mathcal{S}_{v_{r+1}, \dots, v_n}$. In addition, $\bar{\mathbf{Y}}^T$ is r -rank. Thus (6) means that $\mathcal{S}_{v_{r+1}, \dots, v_n}$ is the kernel space of $\bar{\mathbf{Y}}^T$ which essentially spanned by $\bar{v}_{r+1}, \dots, \bar{v}_n$. This completes the proof. \square

Theorem 1 implies that if Ψ does not cause any one of v_1, \dots, v_r to change with one of v_{r+1}, \dots, v_n , then Ψ has no influence on making $\mathcal{S}_{v_{r+1}, \dots, v_n}$ deviated from $\mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$. In this case, $\mathcal{S}_{v_{r+1}, \dots, v_n}$ is the ground truth of the kernel space of $\bar{\mathbf{Y}}^T$. When Ψ is a random matrix, how possible we can get $\mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$ from $\mathcal{S}_{v_{r+1}, \dots, v_n}$ refers to the following theorem.

Theorem 2: If all entries of Ψ are independent random variables, each with 0 mean and bounded by $[-\Psi_{bnd}, \Psi_{bnd}]$, then the probability for holding $\mathcal{S}_{v_{r+1}, \dots, v_n} = \mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$ satisfies

$$\mathcal{P}(\mathcal{S}_{v_{r+1}, \dots, v_n} = \mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}) \geq \prod_{i=1}^r \prod_{j=r+1}^n \left(1 - B_E \left(\frac{\sigma_i^2 - \sigma_j^2}{2(\sigma_i^2 + \sigma_j^2)\Psi_{bnd}} \right) \right) \quad (10)$$

where $B_E(x)$ is the Eaton's bound function.

Proof: The order of v_1, \dots, v_n is due to the order of $\sigma_1 \geq \dots \geq \sigma_n$. The projections of all column vectors of \mathbf{Y}^T onto $\mathcal{S}_{v_1, \dots, v_n}$ are

$$\mathbf{P}_{\mathcal{S}_{v_1, \dots, v_n}}(\mathbf{Y}^T) = [v_1, \dots, v_n] \begin{bmatrix} \sigma_1 u_1^T \\ \vdots \\ \sigma_n u_n^T \end{bmatrix} \quad (11)$$

which demonstrates that the energy of projecting all column vectors of \mathbf{Y}^T onto v_i is not less than that onto v_{i+1} thanks to $\|\sigma_1 u_1\| \geq \dots \geq \|\sigma_n u_n\|$. The vectors v_1, \dots, v_n form an orthogonal frame in R^n . If the energy values of projecting all column vectors of the ground truth matrix $\bar{\mathbf{Y}}^T$ onto v_{r+1}, \dots, v_n each are less than that on v_1, \dots, v_r each, we can get $\mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$ from \mathbf{Y} because in this case $\mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n} = \mathcal{S}_{v_{r+1}, \dots, v_n}$ totally holds according to Theorem 1. To discuss the influence of Ψ on the distribution of the energy values of projecting all column vectors of \mathbf{Y}^T onto v_1, \dots, v_n , we project the column vectors of Ψ^T onto the fame formed by v_1, \dots, v_n in $\mathcal{S}_{v_1, \dots, v_n}$.

$$\mathbf{P}_{\mathcal{S}_{v_1, \dots, v_n}}(\Psi^T) = [v_1, \dots, v_n] \begin{bmatrix} v_1^T \Psi^T \\ \vdots \\ v_n^T \Psi^T \end{bmatrix}. \quad (12)$$

Combining (11) and (12) tells that the energy values of projecting all column vectors of $\bar{\mathbf{Y}}^T$ onto v_i are as follows

$$\|\sigma_i u_i^T - v_i^T \Psi^T\|^2 = \|v_i^T \Psi^T\|^2 + \sigma_i^2 - 2\sigma_i u_i^T \Psi v_i. \quad (13)$$

The entries of Ψ^T are real independent random variables. In terms of the rotational invariance property of Gaussian distributions [18], the following equations

$$\|v_1^T \Psi^T\|^2 = \dots = \|v_n^T \Psi^T\|^2 \quad (14)$$

hold statistically.

If $\|\sigma_i u_i^T - v_i^T \Psi^T\|^2$ for $1 \leq i \leq r$ are larger than $\|\sigma_j u_j^T - v_j^T \Psi^T\|^2$ for $r+1 \leq j \leq n$, then $\mathcal{S}_{v_{r+1}, \dots, v_n} = \mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$ holds. Otherwise, there exists $\mathcal{S}_{v_{r+1}, \dots, v_n} \neq \mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$ and at least one of the following inequalities holds based on (13) and (14)

$$\sigma_i^2 - 2\sigma_i u_i^T \Psi v_i < \sigma_j^2 - 2\sigma_j u_j^T \Psi v_j, \quad (15)$$

which equals to

$$\begin{aligned} & \sigma_i u_i^T \Psi v_i - \sigma_j u_j^T \Psi v_j \\ &= (\sigma_i v_i^T \otimes u_i^T - \sigma_j v_j^T \otimes u_j^T) \text{vec}(\Psi) > \frac{\sigma_i^2 - \sigma_j^2}{2}. \end{aligned} \quad (16)$$

Due to

$$\begin{aligned} & \|\sigma_i v_i^T \otimes u_i^T - \sigma_j v_j^T \otimes u_j^T\|_2^2 \\ &= (\sigma_i v_i^T \otimes u_i^T - \sigma_j v_j^T \otimes u_j^T) (\sigma_i v_i \otimes u_i - \sigma_j v_j \otimes u_j) \\ &= \sigma_i^2 + \sigma_j^2, \end{aligned}$$

the equation (16) can be equivalently changed into

$$\frac{\sigma_i v_i^T \otimes u_i^T - \sigma_j v_j^T \otimes u_j^T}{\sigma_i^2 + \sigma_j^2} \text{vec}(\Psi) > \frac{\sigma_i^2 - \sigma_j^2}{2(\sigma_i^2 + \sigma_j^2) \Psi_{bnd}}. \quad (17)$$

Based on (17), using Eaton's inequality we get the upper bound probability that one relation in (15) holds

$$\begin{aligned} & \mathcal{P}(\sigma_i^2 - 2\sigma_i u_i^T \Psi v_i < \sigma_j^2 - 2\sigma_j u_j^T \Psi v_j) \\ & < B_E \left(\frac{\sigma_i^2 - \sigma_j^2}{2(\sigma_i^2 + \sigma_j^2) \Psi_{bnd}} \right). \end{aligned} \quad (18)$$

For any fixed $i \in \{1, \dots, r\}$, when anyone relation in (15) holds, there is $\mathcal{S}_{v_i} \neq \mathcal{S}_{\bar{v}_i}$. Thus, based on (18) the lower bound probability that $\mathcal{S}_{v_i} = \mathcal{S}_{\bar{v}_i}$ is

$$\mathcal{P}(\mathcal{S}_{v_i} = \mathcal{S}_{\bar{v}_i}) \geq \prod_{j=r+1}^n \left(1 - B_E \left(\frac{\sigma_i^2 - \sigma_j^2}{2(\sigma_i^2 + \sigma_j^2) \Psi_{bnd}} \right) \right). \quad (19)$$

So the lower bound probability that $\mathcal{S}_{v_{r+1}, \dots, v_n} = \mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$ is as given in (10). This completes the proof. \square

Remark 1: The Eaton's bound function $B_E(x)$ is monotonically decreasing. For a given $i \in \{1, \dots, r\}$, $B_E \left(\frac{\sigma_i^2 - \sigma_j^2}{2(\sigma_i^2 + \sigma_j^2) \Psi_{bnd}} \right)$ usually decreases with $j \in \{r+1, \dots, n\}$ due to $\sigma_{r+1} \geq \sigma_{r+2} \geq \dots \geq \sigma_n$, thus (18) indicates that removing Ψ is less possible to make the energy projected by all columns of \mathbf{Y}^T onto v_i less than that onto v_j with larger j . That is to say, for larger j , Ψ has smaller influence on causing v_j to get away from $\mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$, and accordingly $v_j \in \mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$ will hold in higher probability.

Remark 2: Theorem 2 shows that smaller Ψ_{bnd} makes the lower bound of $\mathcal{P}(\mathcal{S}_{v_{r+1}, \dots, v_n} = \mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n})$ larger. That is to say, noise matrix Ψ with smaller Ψ_{bnd} has less influence on $\mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$. Meanwhile, Theorem 2 also shows that if σ_i^2 is much larger than σ_j^2 , the lower bound of $\mathcal{P}(\mathcal{S}_{v_{r+1}, \dots, v_n} = \mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n})$ is also larger, meaning that larger $\sigma_i^2 - \sigma_j^2$ for $1 \leq i \leq r$ and $r+1 \leq j \leq n$ make $\mathcal{S}_{v_{r+1}, \dots, v_n}$ more prone to $\mathcal{S}_{\bar{v}_{r+1}, \dots, \bar{v}_n}$.

The rank of $\bar{\mathbf{Y}}$ is r , and any submatrix randomly chosen from $\bar{\mathbf{Y}}$, $\bar{\mathbf{Y}}_{\langle k \rangle \times \langle l \rangle}$ with $k, l \geq r$, $\langle k \rangle \subset [m]$ and $\langle l \rangle \subset [n]$, usually has rank not larger than r . Thus the right singular vectors corresponding to the $l-r$ smallest singular values ought to be the null vectors of $\bar{\mathbf{Y}}_{\langle k \rangle \times \langle l \rangle}$. In terms of Remark 1, the subspace spanned by the right singular vectors corresponding to small or trivial singular values of $\mathbf{Y}_{\langle k \rangle \times \langle l \rangle}$ is close to that of $\bar{\mathbf{Y}}_{\langle k \rangle \times \langle l \rangle}$. Thus we can use the right singular vectors corresponding to the $\ell \in [1, l-r]$ smallest singular values of $\mathbf{Y}_{\langle k \rangle \times \langle l \rangle}$ to restrict \mathbf{V} . Under the assumption that known entries are homogeneously distributed, we can use the following two methods to randomly extract submatrices from \mathbf{Y} . The submatrices are with size $m\rho^l \times l$ or $k \times n\rho^k$, where $m\rho^l$ and $n\rho^k$ are two modes with k, l definitely predefined; that is to say, when randomly choosing l columns or k rows from \mathbf{Y} , the valid row number or the valid column number will swing about $m\rho^l$ or $n\rho^k$.

M1: randomly choose l columns usually with $l = r+1$ and take all the rows whose entries at the chosen columns are all known, and get a matrix with statistical size $m\rho^l \times l$;

M2: operation like 1) in horizontal, first randomly choose k rows, then select columns accordingly, and get a matrix with size $k \times n\rho^k$ statistically.

We do not know which entries in \mathbf{Y} is more severely disturbed, and each known entry is valuable and should be used as possible as we can. To visit as many known entries in \mathbf{Y} as possible, we need extract many submatrices $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$, $\mathbf{Y}_{\langle m\rho^l \rangle \times \langle l \rangle}$ or their combination. For the simplicity of discussion, assume we choose a submatrix $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ in each trial. Then how many trials we need to make all known entries each visited in a special probability? About this question we present Theorem 3 after introducing Lemma 1.

Lemma 1: Let $\varphi(x, \rho) = x \log \frac{x}{\rho} + (1-x) \log \frac{1-x}{1-\rho}$,

$\phi(x)$ denote the probability density function of standard normal distribution and $x_{n,\rho}$ be a binomial random variable: $\mathcal{P}(x_{n,\rho} < k) = \sum_{i=0}^k \binom{n}{i} \rho^i (1-\rho)^{n-i}$. An increasing sequence $\{\mathcal{C}_{n,\rho}(k)\}_{k=0}^n$ is defined as $\mathcal{C}_{n,\rho}(0) \equiv (1-\rho)^n$, $\mathcal{C}_{n,\rho}(n) \equiv 1 - \rho^n$ and $\mathcal{C}_{n,\rho}(k) \equiv \int_{-\infty}^{\text{sgn}(kn-1-\rho)\sqrt{2n\varphi(kn-1,\rho)}} \phi(x) dx$ for $1 \leq k < n$. Then

$$\mathcal{C}_{n,\rho}(k) \leq \mathcal{P}(x_{n,\rho} < k) \leq \mathcal{C}_{n,\rho}(k+1) \quad (20)$$

and equalities hold only for $k=0$ or $k=n-1$ [19].

Theorem 3: If known entries are homogeneously distributed in \mathbf{Y} with density ρ (as defined in (3)) and in each trial submatrix $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ is randomly extracted with $k > r$, then at most

$$\mathcal{I} \leq \frac{mn\rho}{k(r+1)(1-\mathcal{C}_{n,\rho^k}(r+2))} \ln \frac{1}{1-\epsilon} \quad (21)$$

trials are required to make each known entry of \mathbf{Y} visited with the probability at least ϵ .

Proof: Let $\mathcal{P}(\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle})$ denote the probability that a known entry will be visited in a trial. The column number $n\rho^k$ is a mode, and in each trial the column number, say i , may be any number from 0 to n . The probability that $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ has i columns is $\binom{n}{i} (\rho^k)^i (1-\rho^k)^{n-i}$ for $i = 0, \dots, n$. If $i > r$ (because we only choose the submatrix satisfying $k, i > r$), $k \times i$ known entries are visited, and each known entry is visited with probability $\frac{ki}{mn\rho}$ in this case. So

$$\begin{aligned} \mathcal{P}(\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}) &= \sum_{i=r+1}^n \binom{n}{i} (\rho^k)^i (1-\rho^k)^{n-i} \frac{ki}{mn\rho} \\ &\stackrel{1)}{\geq} \frac{k(r+1)}{mn\rho} (1-\mathcal{C}_{n,\rho^k}(r+2)) \end{aligned} \quad (22)$$

where we have used Lemma 1 in 1).

In \mathcal{I} trials, the probability that a known entry of \mathbf{Y} will be visited is $1 - (1 - \mathcal{P}(\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}))^{\mathcal{I}}$, by which we get $1 - \mathcal{P}(\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle})^{\mathcal{I}} < 1 - \epsilon$ for the requirement of visiting each known entry with the probability at least ϵ . Finally we get

$$\mathcal{I} \leq \frac{1}{\mathcal{P}(\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle})} \ln \left(\frac{1}{1-\epsilon} \right) \quad (23)$$

where the known inequality $\ln(1-x) \leq -x$ for $x \in (0, 1)$ has been used. From (22) and (23), (21) is derived. This completes the proof.

If much time is spent finding enough $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ to make each known entry visited with probability at least ϵ , efficiency becomes low. The upper bound in Theorem 3 tells how r , k and ρ affect the trial number.

Remark 3: For a given \mathbf{Y} and ϵ , the upper bound of \mathcal{I} increases with ϵ , and is proportional to $\frac{\rho}{k(r+1)(1-\mathcal{C}_{n,\rho^k}(r+2))}$ which indicates that larger ρ , k or r does not mean larger bound as $\mathcal{C}_{n,\rho^k}(r+2)$ is increasing with k and r and decreasing with ρ .

From each $\mathbf{Y}_{\langle m\rho^l \rangle \times \langle l \rangle}$ or $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ with row and column numbers larger than r , we can get $l-r$ or $n\rho^k-r$ right singular vectors (corresponding to the $l-r$ or $n\rho^k-r$ smallest singular values) accordingly, denoted as ζ_j for $j = 1, \dots, l-r$ or for $j = 1, \dots, n\rho^k-r$. To constrain

\mathbf{V} , we can use all ζ_j , or choose some ζ_ℓ corresponding to $\ell \leq l-r$ or $\ell \leq n\rho^k-r$ smallest singular values in terms of Remark 1. Especially, when $n\rho^k > k > r$ or $l > m\rho^l > r$ holds we can choose the null vectors.

We can extend ζ_j from l or $n\rho^k$ to n dimension, denoted as

$$\xi_j = e(\zeta_j) \in R^n, \quad (24)$$

by substituting the l or $n\rho^k$ places of an n -dimensional zero vector (corresponding to the places where the columns of \mathbf{Y} are chosen) with the entries of ζ_j accordingly. Remark 1 tells that ξ_j corresponding to smaller singular values of $\mathbf{Y}_{\langle m\rho^l \rangle \times \langle l \rangle}$ or $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ has higher probability that ξ_j belongs to the null space of $\bar{\mathbf{Y}}$. So we use the special ξ_j as the null vector of $\bar{\mathbf{Y}}$, and further get

$$\bar{\mathbf{V}}^T \xi_j = 0 \quad (25)$$

where $\bar{\mathbf{V}}$ can be seen as the optimum of \mathbf{V} in (2).

The equations (24) and (25) tell that all the vectors ξ_j resulted from the same $\mathbf{Y}_{\langle m\rho^l \rangle \times \langle l \rangle}$ or $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ only constrain the same rows of $\bar{\mathbf{V}}$. To constrain all rows of $\bar{\mathbf{V}}$, we need to randomly extract $\mathbf{Y}_{\langle m\rho^l \rangle \times \langle l \rangle}$ or $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ for many times, say \mathcal{I} , and ξ_j^i is additionally indexed by i for $1 \leq i \leq \mathcal{I}$. All ξ_j^i can be organized as a matrix

$$\Xi = [\xi_j^i] \in R^{n \times z} \quad (26)$$

where z is dependent on \mathcal{I} and on how many ξ_j vectors are obtained from $\mathbf{Y}_{\langle m\rho^l \rangle \times \langle l \rangle}$ or $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ each. If the rank of Ξ exceeds $n-r$, using

$$\hat{\mathbf{V}} = \arg \min_{\mathbf{V}} \|\mathbf{V}^T \Xi\|, \text{ s.t. } \|\mathbf{V}\| = 1, \quad (27)$$

we can get the optimized \mathbf{V} as $\hat{\mathbf{V}}$, which is close to or amounts to $\bar{\mathbf{V}}$. The goal of the constraints $\|\mathbf{V}\| = 1$ in (27) is to prevent $\hat{\mathbf{V}}$ from becoming trivial in the optimization procedure, and the constraints can be replaced by many other forms.

4 THE ALGORITHM

In this section, the algorithm is described, followed by the analysis on its computational cost and memory requirements.

4.1 Description of the Algorithm

In view of (27), we need to construct Ξ first. Each column vector of Ξ is calculated from a randomly chosen submatrix $\mathbf{Y}_{\langle m\rho^l \rangle \times \langle l \rangle}$ or $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$. For given $k (> r)$ rows of \mathbf{Y} , it is critical for efficiency to quickly choose the columns whose entries at the chosen rows are known. Only when the number of the chosen columns is larger than r , can a submatrix be formed. In this procedure, only logical comparisons are operated on \mathbf{W} which only has boolean entries, thus extracting a submatrix $\mathbf{Y}_{\langle k \rangle \times \langle n\rho^k \rangle}$ is usually very fast. If we keep k or l constant in all trials, the concrete value of k or l is related to the distribution of known entries, and larger k or l

will make $n\rho^k$ or $m\rho^l$ less. So, taking $k, l = r + 1$ is usually feasible. In each trial, the submatrix is randomly extracted, and the submatrices chosen in different trials have no relations to each other. Thus randomly extracting submatrices can be implemented in parallel and by multiprocessor architectures as illustrated in [2]. In practice, how many submatrices are required can refer to Theorem 3.

In implementing (27), many concrete forms can be adopted such as the following quadratic form

$$\hat{\mathbf{V}} = \arg \min_{\mathbf{V}} \|\mathbf{V}^T \Xi\|_2, \text{ s.t. } \|\mathbf{V}\|_2 = 1. \quad (28)$$

In this case $\hat{\mathbf{V}}$ can take the r left singular vectors corresponding to the r smallest singular values of Ξ . Actually, in this case storing all entries of Ξ is unnecessary, and it only needs to store n^2 real values of $\sum_i \sum_j \xi_j^i (\xi_j^i)^T = \Xi \Xi^T$. A parallel way to quickly solve (28) is through dynamic computation [20]. An alternative approach of (27) is

$$\hat{\mathbf{V}} = \arg \min_{\mathbf{V}} \|\mathbf{V}^T \Xi\|_1 \quad (29)$$

with the regularization constraints which cannot let the optimum of (29) be zeros. There are so many norm definitions, thus (27) has many other concrete forms. After getting $\hat{\mathbf{V}}$, based on (2) we can work out $\hat{\mathbf{U}}$ as follows

$$\hat{\mathbf{U}} = \min_{\mathbf{U}} \|\mathbf{W}_{m \times n} \odot (\mathbf{Y}_{m \times n} - \mathbf{U}_{m \times r} \hat{\mathbf{V}}_{n \times r}^T)\|. \quad (30)$$

The product $\hat{\mathbf{U}} \hat{\mathbf{V}}^T$ is the low-rank decomposition of \mathbf{Y} .

The Eqs. (27) and (30) provide the two fundamental formulas calculating the low-rank decomposition of \mathbf{Y} , and each of them can be implemented in quadratic programming and be solved in polynomial time. In contrast, the primary problem (2) is indefinite, and is NP-hard even when taking L_2 norm. To make (2) solvable in polynomial time, with Ξ as the inter-medium, (2) is transformed into Eqs. (27) and (30). In summary, to optimize (2) the proposed algorithm is implemented via the following three steps:

- step 1: In terms of a prior knowledge, the row number $k > r$ or the column number $l > r$ is specified, and further \mathcal{I} is fixed with a preconditioned probability ϵ for visiting known entries based on Theorem 3. In practice, it is feasible to evaluate k or l with $r + 1$.
- step 2: By **M1** or **M2**, randomly choose submatrix. If the row and column numbers of the submatrix are not less than r , calculate ξ_j^i corresponding to the null or small singular values and save them in Ξ . Repeat I trials, and Ξ is constructed finally.
- step 3: The Eq. (27) provides a framework to work out $\hat{\mathbf{V}}$, and (28) or (29) can be adopted instead. By (30), $\hat{\mathbf{U}}$ is solved and then the low-rank decomposition of \mathbf{Y} is calculated as $\hat{\mathbf{U}} \hat{\mathbf{V}}^T$. When the norm other than Frobenius norm is adopted, the recurrent dynamic system as used in [5] can be used to solve (27) and (30) when L_1 norm is adopted.

4.2 CPU Time and Memory Requirements

Assume **M2** is used in each trial, and only the submatrix $\mathbf{Y}_{\langle k \rangle \times \langle n \rho^k \rangle}$ having null vectors is chosen. For each submatrix, calculating its $n\rho^k - k$ null vectors ξ_j^i costs $kn\rho^k(n\rho^k - k)$ flops using the naive SVD, and calculating $\sum_{j=1}^{n\rho^k - k} \xi_j^i (\xi_j^i)^T$ needs $(n\rho^k - k)(n\rho^k)^2$ flops. Thus calculating $\Xi \Xi^T$ (applicable to using L_2 norm as used in (28)) using **M2** costs $[kn\rho^k(n\rho^k - k) + (n\rho^k - k)(n\rho^k)^2] = \mathcal{O}(n^3 \rho^{3k})$ flops due to $n\rho^k > k$. Similarly, calculating $\Xi \Xi^T$ using **M1** costs $[m\rho^l l(l - r) + (l - r)l^2] = \mathcal{O}(ml^2 \rho^l)$ flops.

When using (28) to solve $\hat{\mathbf{V}}$, we only need to calculate r left singular vectors of $\Xi \Xi^T$ corresponding to the r smallest singular values, and this step costs rn^2 flops when solving eigen-pairs with Lanczos technique and the homotopy method. Solving each row of $\hat{\mathbf{U}}$ by (30) costs $r^2 n \rho^2 + r^2 \log(r) + r^2$ flops if L_2 norm is used. In total, **RSM** needs $\mathcal{O}(mr^2 \rho^r)$ or $\mathcal{O}(n^3 \rho^{3r})$ plus $m(r^2 n \rho^2 + r^2 \log(r) + r^2) + rn^2$ flops since l and k take $r + 1$ usually. Actually, calculating $\hat{\mathbf{V}}$ from $\Xi \Xi^T$ and accordingly calculating $\hat{\mathbf{U}}$ only run once, and the procedures are very fast in scientific computation platforms such as Matlab and the computation load can be neglected. So **RSM** almost costs $\mathcal{O}(mr^2 \rho^r)$ or $\mathcal{O}(n^3 \rho^{3r})$, as shown in (4). The total trial number \mathcal{I} is bound by Theorem 3, and actually the theory proposed in [1] seems to indicate that when \mathcal{I} is much less than the bound, the precision of the proposed algorithm is also very competitive, which is confirmed by the following experimental results. In contrast, the algorithm in [3] requires $\mathcal{O}(nmr + r^2(m + n))$ flops in each iteration. So our algorithm is computation-saving compared to the state-of-the-art algorithms such as the one in [3].

In performing low-rank decomposition of $\mathbf{Y} \in R^{m \times n}$, the algorithms in [16], [17] and [3] need memory spaces to store $\max(mn, 2(m + n)r)$, $m^2 n \rho + mn^2 r \rho + mn \rho + 2mn \rho r + mr^2$ and $mn + (m + n)r$ real numbers, respectively. So, the algorithm in [17] is most memory-consuming, and is not applicable to large-scale matrices \mathbf{Y} . If L_2 norm is adopted in (27) and (30), **RSM** only needs memory space to save $\max(n^2, mr + nr)$ real numbers, where n^2 is resulted from saving $\Xi \Xi^T$ and the term $(m + n)r$ is due to saving $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$. Comparing the memory space requirements of the mentioned algorithms with the proposed algorithm demonstrates that **RSM** is very memory-saving.

5 NUMERICAL RESULTS

In this section, several numerical tests on synthetic and real data were done with the comparison of **RSM** with the state-of-the-art algorithms [3] [16]¹ [17]². All algorithms were implemented and run in Matlab in double precision arithmetic on a PC with one core of a 3.2 GHz Intel Core i5-3470 microprocessor and with 8 GB RAM.

1. <http://www.stanford.edu/~raghuram/optspace/code.html>, accessed at 3/22/2012

2. <http://www.vision.is.tohoku.ac.jp/us/download/>

5.1 Synthetic Data Tests

First we use synthetic data to test the algorithms, and in view of (5) the synthetic random matrices \mathbf{Y} and \mathbf{W} are produced as follows

$$\begin{aligned}\bar{\mathbf{Y}} &= \text{randn}(m, r) \times \text{randn}(r, n), \Psi = \sigma \times \text{randn}(m, n), \\ \mathbf{W} &= \text{rand}(m, n) \leq \rho\end{aligned}\quad (31)$$

where $\text{randn}(n, r)$ or $\text{rand}(n, r)$ denotes an n -by- r matrix whose each entry is a pseudorandom value drawn from the standard normal distribution or from the standard uniform distribution on the open interval $(0, 1)$, and $\sigma > 0$ calibrates the spectrum of the noise. In (31), $\bar{\mathbf{Y}} = \text{randn}(m, r) \times \text{randn}(r, n)$ constructs the ground truth of a r -rank matrix, and Ψ is the white noise matrix with power spectrum σ for each entry. The $\rho \in (0, 1)$ can refer to (3), and the matrix \mathbf{W} built in (31) indicates that known entries are homogeneously distributed. The L_2 error is calculated as follows

$$e = (\sum_{i,j} w_{ij})^{-0.5} \|\mathbf{W} \odot (\mathbf{Y} - \hat{\mathbf{U}}\hat{\mathbf{V}}^T)\|_F \quad (32)$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Let $r = 3$ and \mathcal{I} takes $15n \sim 35n$. In each trial, $\mathbf{M1}$ is used to randomly extract a submatrix of \mathbf{Y} . Let $l = r + 1$, so each trial can produce $l - r = 1$ vector to restrict \mathbf{V} , statistically. Table 1 lists experimental results for different combinations of m, n, ρ and σ along with the comparison of **RSM** with the state of the arts. The algorithm in [17] is very memory-consuming, and it does not work for the synthetic matrices due to memory overflow, so it does not join tests. The algorithm in [3] contains inter and outer loops, and the two iteration numbers are evaluated with 10 as each iteration seems time-consuming. The algorithm in [16] has one loop, and its iteration number takes 50 as each iteration is too much time-consuming. The CPU time comparison of **RSM** with that of [3] and [16] is also given in Table 1 with titles $t_{[3]}/t$ and $t_{[16]}/t$.

From Table 1, we can observe the following points which are also consistent with the results of more extensive experimentation performed by the author.

- 1) For all random matrices \mathbf{Y} , the error differences between **RSM** and the algorithm in [3], $|e - e_{[3]}|$, are almost zero except one is $2\text{E-}4$ and the other is $1\text{E-}4$. So, the low-rank decomposition precisions of the two algorithms are almost the same. However, $t_{[3]}/t$ indicates that **RSM** is $4.97 \sim 12.00$ times faster the state-of-the-art algorithm. Compared with **RSM**, the algorithm in [16] has low precision due to $e < e_{[16]}$ and is $16 \sim 46$ times slower than **RSM** in terms of $t_{[16]}/t$.
- 2) When only row number m or column number n increases, the CPU time increases accordingly. To discuss the increasing speed of CPU time with m and n , we plot the points with m or n as the x-coordinate and with CPU time as the y-coordinate. All the values of m, n and CPU time are divided by their corresponding minimal values in order to clearly show the relations of CPU time vs m or n

while to remove the influence of starting points. We use linear relations to fit in with the points, as shown in Fig. 1, from which we can observe that, the slope of the linear relation corresponding to the algorithms in [3], [16] or **RSM** is larger than, almost equals to, or is less than 1, respectively. The facts show that, the CPU time of **RSM** increases most slowly with respect to m or n among the three algorithms.

- 3) All algorithms need more CPU time with larger ρ , and spectrum of noise σ appears to have no influence on CPU time. On the whole, the errors of **RSM** and the algorithm in [3] are less than σ while sometimes that of the algorithm in [16] are not.

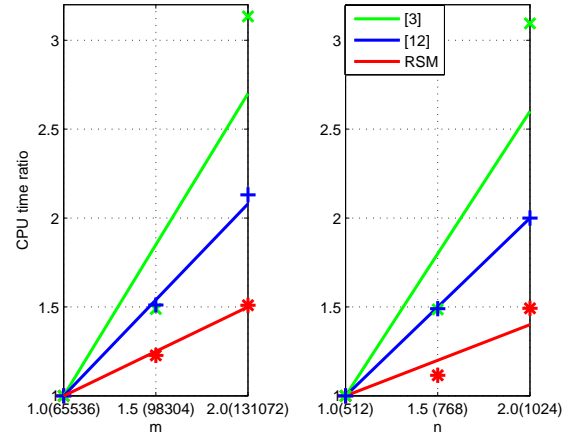


Fig. 1. The linear fitting relations of CPU time vs m or n . Comparing the relations with each other demonstrates that the increasing ratio of **RSM** is minimal.

5.2 Real Data Tests

The real data sets consist of three image sequences: dinosaur, giraffe and face¹, whose m, n, r and ρ are listed in Table 2. The dinosaur sequence consists of 36 images with the resolution of 720×576 that are taken from an artificial dinosaur on a turntable, and frame 13 of the sequence is shown as the left image of Fig. 2. The giraffe sequence contains 120 frames (the 47th one is shown as the middle image of Fig. 2), and its measurement matrix is about the occluded motion of a giraffe. The face sequence demonstrates a static face lit by a distant light source from different directions, and the 10th frame refers to the right image of Fig. 2.

TABLE 2
The details of the real data sets

Datasets	m	n	r	ρ
giraffe	240	167	6	69.8%
dinosaur	319	72	4	28%
face	2944	20	4	58%

1. <http://www.robots.ox.ac.uk/~abm/>

TABLE 1
Test and comparison on synthetic random matrix

matrix parameters				RSM		[3]		[16]		time comparison	
m	n	ρ	σ	t	e	$t_{[3]}$	$e_{[3]}$	$t_{[16]}$	$e_{[16]}$	$t_{[3]}/t$	$t_{[16]}/t$
65536	1024	0.2	0.2	5.24E1	0.1987	3.13E2	0.1985	0.92E3	0.2005	5.97	17.56
98304	1024	0.2	0.2	6.43E1	0.1986	4.66E2	0.1985	1.39E3	0.1988	7.24	21.62
131072	1024	0.2	0.2	7.91E1	0.1985	9.81E2	0.1985	1.96E3	0.2180	12.40	24.78
131072	768	0.2	0.2	5.91E1	0.1981	4.71E2	0.1981	1.46E3	0.2452	7.97	24.70
131072	512	0.2	0.2	5.30E1	0.1971	3.17E2	0.1971	0.98E3	0.3019	5.98	18.49
131072	1024	0.5	0.1	8.42E1	0.0997	8.49E2	0.0997	3.96E3	0.1081	10.08	47.03
131072	1024	0.4	0.1	7.57E1	0.0996	8.08E2	0.0996	2.83E3	0.1158	10.67	37.38
131072	1024	0.3	0.1	5.91E1	0.0995	7.68E2	0.0995	2.24E3	0.1000	13.00	37.90
131072	1024	0.3	0.3	5.95E1	0.2985	7.60E2	0.2985	2.16E3	0.3004	12.77	36.30
131072	1024	0.3	0.5	6.04E1	0.4976	7.73E2	0.4975	2.20E3	0.4993	12.80	36.42



Fig. 2. From left to right, frames 13, 47 and 10 of the three image sequences: dinosaur, giraffe and face, respectively.

To ensure the low-rank approximation precision, both inter and outer iteration numbers of the algorithm in [3] take 1000, and the single iteration number of the algorithm in [16] takes 10000. Meanwhile, the sizes of the real data sets are small, and memory overflow does not arise for the algorithm in [17], so it also joins tests, and its iteration number takes 1000. The experimental results are listed in Table 3, from which the following points we can observe.

- 1) The CPU time comparison value ranges from 8.96 to 89.60, meaning that **RSM** is $7.96 \sim 88.96$ times faster than the comparing methods. The algorithm in [3] is more time-consuming than that in [16], which is also more time-consuming than that in [17].
- 2) The precision of **RSM** is close to that of [3], better than that of [16] and inferior to that in [17]. Through the algorithm in [17] has good precision, both theoretical analysis and the experimentation on synthetic data sets show that the algorithm is not applicable to large-scale data sets because it is rather memory-consuming. Each point on the artificial dinosaur rotates with the turntable and forms a circle in 3D Euclidean space. After projective transformation of the camera, the circle becomes into an ellipse. So the recovered tracks of the dinosaur sequence ought to be closed and smooth ellipses. The recovered tracks of the dinosaur sequence are shown in Fig. 3, from which we can see the tracks recovered by **RSM** is closer to circles than all the other comparing algorithms though the error of **RSM**, $e = 1.1826$, is a little larger than that of L2-Wiberg, $e_{[17]} = 1.0847$.
- 3) Combining Table 2 and Table 3, we can see that

$t_{[3]}/t$, $t_{[16]}/t$ and $t_{[17]}/t$ corresponding to giraffe are the minimal among the three data sets. This fact is in accordance with the computational complexity analysis as given in Section 4.2: comparing (4) with (1) shows that for matrix \mathbf{Y} with $m \gg n$, **RSM** will have better efficiency. Of course, for giraffe data, the minimal value of $t_{[3]}/t$, $t_{[16]}/t$ and $t_{[17]}/t$ is 9.46, which shows **RSM** is still much faster than the comparing methods even for matrices without $m \gg n$ such as giraffe data.

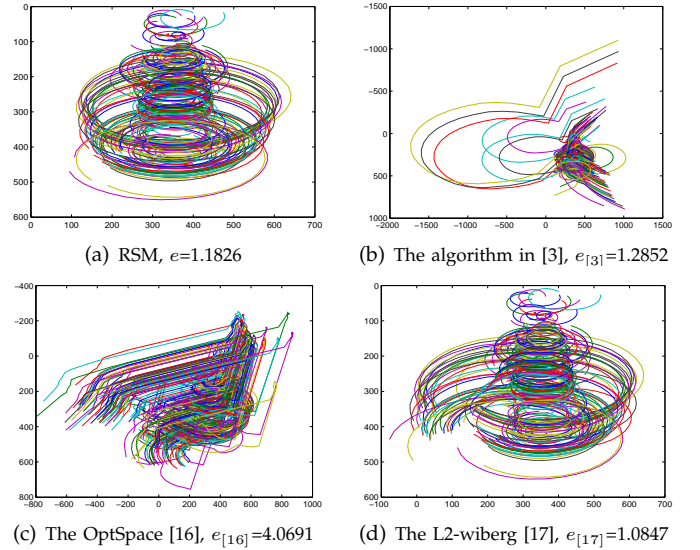


Fig. 3. The tracks and errors gotten by **RSM** and the state of the art for the dinosaur sequence. The true tracks are ellipses because each point onto the artificial dinosaur on the turntable forms a circle, and circle becomes ellipse due to camera projection. Comparing the tracks with each other tells that the tracks recovered by **RSM** are most approximate to closed ellipses

6 CONCLUSIONS AND FUTURES

In this paper a Random SubMatrix framework (**RSM**) has been proposed for calculating the low-rank decomposition of matrix, $\mathbf{Y}_{m \times n}$, with known entry percentage ρ . **RSM** uses submatrices randomly chosen from $\mathbf{Y}_{m \times n}$

TABLE 3
Tests and comparisons on real data sets

data	RSM		[3]		[16]		[17]		time comparison		
	t	e	$t_{[3]}$	$e_{[3]}$	$t_{[16]}$	$e_{[16]}$	$t_{[17]}$	$e_{[17]}$	$t_{[3]}/t$	$t_{[16]}/t$	$t_{[17]}/t$
dinosaur	1.73	1.1826	1.55E2	1.2852	7.29E1	4.0691	1.55E1	1.0847	89.60	42.14	8.96
giraffe	5.73	0.3833	2.47E2	0.3344	1.19E2	0.9431	5.42E1	0.3228	43.11	20.77	9.46
face	6.01	0.0239	3.55E2	0.0225	1.40E2	0.0361	1.43E2	0.0226	59.07	23.29	23.79

to get the low-rank approximation of $\mathbf{Y}_{m \times n}$, $\hat{\mathbf{U}}_{m \times r} \hat{\mathbf{V}}_{n \times r}^T$. All entries of each submatrix are known. Due to the just proved theorem that noises are less easy to make the singular vector space corresponding to smaller or trivial singular values to change with the subspace corresponding to anyone of the r largest singular values, we can choose the singular vectors corresponding to smaller or trivial singular values of each submatrix to constrain some rows of $\hat{\mathbf{V}}_{n \times r}$. When submatrices are extracted enough to constrain all rows, $\hat{\mathbf{V}}_{n \times r}$ is calculated and accordingly $\hat{\mathbf{U}}_{m \times r}$ is also gotten. Compared with the state-of-the-art algorithms [3] [16] [17] which have complexity $\mathcal{O}(mn r + r^2(m + n))$ or need memory space to save $\max(mn, 2mr + 2nr)$ real numbers, **RSM** is very fast as the computational complexity is only $\mathcal{O}(mr^2\rho^r)$ or $\mathcal{O}(n^3\rho^{3r})$; if the approximation is measured in L_2 norm, **RSM** only needs memory space for saving $\max(n^2, mr + nr)$ real values, so **RSM**, meanwhile, is very memory-saving. These advantages have been verified by experimental results on synthetic and real data sets. On random matrices with different combination of m , n and ρ , such as $\mathbf{Y}_{131072 \times 1024}$, **RSM** is $4.97 \sim 46.03$ times faster than the state-of-the-art algorithms apart from the one [17] always causing memory overflow. On the real data sets, **RSM** is $8.46 \sim 88.60$ times faster. Except the efficiency and memory saving advantages, the low-rank approximation precision of **RSM** is considerably high, and is close to or equivalent to the best of the state-of-the-art algorithms.

Whereas the results of numerical experiments are in reasonably close agreement with theoretical analysis, we find that when random submatrix number is much smaller than the upper bound given in Theorem 3 the low-rank approximation precision is still considerably high. So, tightening the bound given in Theorem 3 is a future task worthwhile to explore. The Equations (27) and (30) only provide a general way for low-rank decomposition, and the metric there can take L_p norm with $p \geq 0$. Especially, when L_1 norm is taken, each column vector of Ξ can take the null singular vectors of each random submatrix, and this operation does not bring about any other additional noises. In this paper the feasibility and the marvelous performance of (27) and (30) are illustrated only using L_2 norm, and how to solve them using other norms are still open. Randomly choosing submatrices can lead to considerable high performance; maybe, combining a prior knowledge of known entry distribution with randomness techniques possibly further benefits precision and computational

speed.

ACKNOWLEDGMENT

This work is supported by NSFC under grants 61173182, 61179071 and 61411130133, and by funding from Sichuan Province (2014HH0048).

REFERENCES

- [1] E. Liberty, F. Woolfe, P.-G. Martinsson, V. Rokhlin, and M. Tytgert, "Randomized algorithms for the low-rank approximation of matrices," *Proceedings of the National Academy of the Sciences of the United States of America*, vol. 104, pp. 20167–20172, December 2007.
- [2] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011.
- [3] R. Cabral, F. D. la Torre, J. ao P. Costeira, and A. Bernardino, "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," *IEEE International Conference on Computer Vision*, pp. 2488–2495, 2013.
- [4] J. Ye, "Generalized low rank approximations of matrices," *Machine Learning*, vol. 61, pp. 167–191, 2005.
- [5] Y. Liu, B. Liu, Y. Pu, X. Chen, and H. Cheng, "Low-rank matrix decomposition in L_1 -norm by dynamic systems," *Image and Vision Computing*, vol. 30, pp. 915–921, November 2012.
- [6] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi, "Practical low-rank matrix approximation under robust l1-norm," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR)*, pp. 1410–1417, 2012.
- [7] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems* (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 2080–2088, 2009.
- [8] P. Chen, "Heteroscedastic low-rank matrix approximation by the Wiberg algorithm," *IEEE Transactions on Signal Processing*, vol. 56, pp. 1429–1439, Apr. 2008.
- [9] J. Liu, S. Chen, Z.-H. Zhou, and X. Tan, "Generalized low-rank approximations of matrices revisited," *IEEE Transactions on Neural Networks*, vol. 21, pp. 621–632, Apr. 2010.
- [10] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, pp. 1548–1566, March 2011.
- [11] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," October 2009.
- [12] A. Eriksson and A. van den Hengel, "Efficient computation of robust weighted low-rank matrix approximations using the l1 norm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1681–1690, September 2012.
- [13] Y. Hu, S. G. Lingala, and M. Jacob, "A fast majorizeminimize algorithm for the recovery of sparse and low-rank matrices," *IEEE Transactions on Image Processing*, vol. 21, pp. 742–753, February 2012.
- [14] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 171–184, January 2013.
- [15] J. Lee, S. Kim, G. Lebanon, and G. Lebanon, "Local low-rank matrix approximation," *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

- [16] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, pp. 2980–2998, June 2010.
- [17] T. Okatani, T. Yoshida, and K. Deguchi, "Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms," *IEEE International Conference on Computer Vision*, pp. 842–849, 2011.
- [18] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 2011.
- [19] A. M. Zubkov and A. A. Serov, "A complete proof of universal inequalities for the distribution function of the binomial law," *Theory of Probability & Its Applications*, vol. 57, no. 3, pp. 539–544, 2013.
- [20] Y. Liu, Z. You, and L. Cao, "A concise functional neural network computing the largest modulus eigenvalues and their corresponding eigenvectors of a real skew matrix," *Theoretical Computer Science*, vol. 367, pp. 273–285, Dec. 2006.